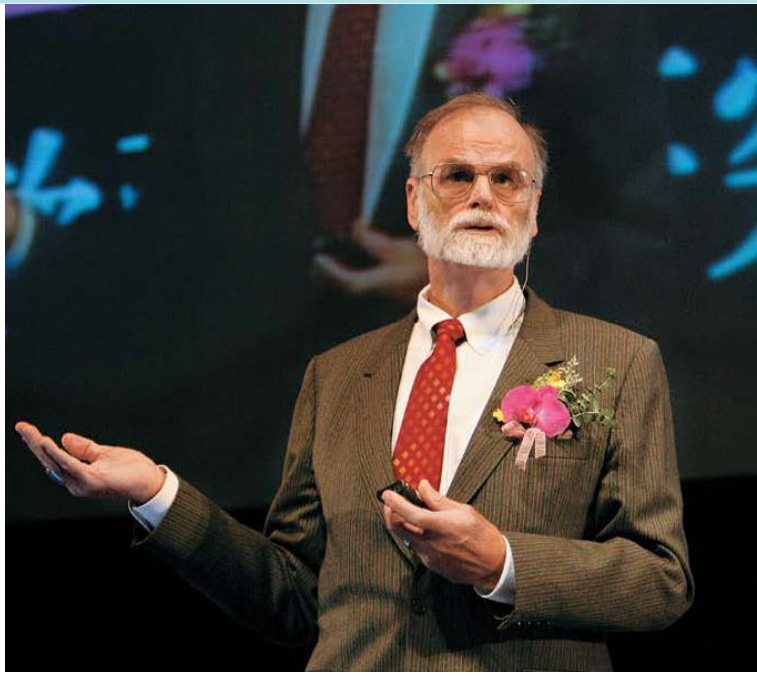
Data Intensive Distributed Computing

Ioan Raicu
Computer Science Department
Illinois Institute of Technology

CS554: Data-Intensive Computing
January 14th, 2015

Paper Discussions

- The Fourth Paradigm
  - Foreward, by Gordon Bell
    - http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_gordon_bell_foreword.pdf
  - Jim Gray on eScience: A Transformed Scientific Method
    - http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_jim_gray_transcript.pdf



Foreward

- Data Intensive Computing ➔ 4th Paradigm
- Scientific research ~ printing press

Data-Intensive Science Components

- Capture
- Data validation through curation
  - finding the right data structures to map into various stores
  - includes the schema and the necessary metadata for longevity and for integration across instruments, experiments, and laboratories
- Analysis
  - workflow pipeline,
  - use of databases (versus a collection of flat files)
  - analysis and modeling
  - data visualization
- Permanent archiving

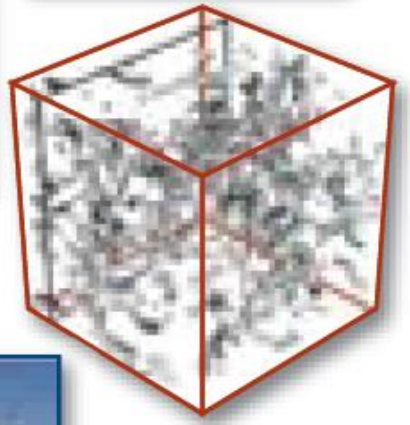
Massive Amounts of Data

- Petabytes of data per day
  - Australian Square Kilometre Array of radio telescopes project
  - CERN's Large Hadron Collider
  - Astronomy's Pan-STARRS5 array of celestial telescopes

Science Paradigms

- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations
- Last few decades:
  a **computational** branch
  simulating complex phenomena
- Today: **data exploration** (eScience)
  unify theory, experiment, and simulation
  - Data captured by instruments or generated by simulator
  - Processed by software
  - Information/knowledge stored in computer
  - Scientist analyzes database/files using data management and statistics

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G p}{3} - \mathrm{K}\frac{c^2}{a^2}$$



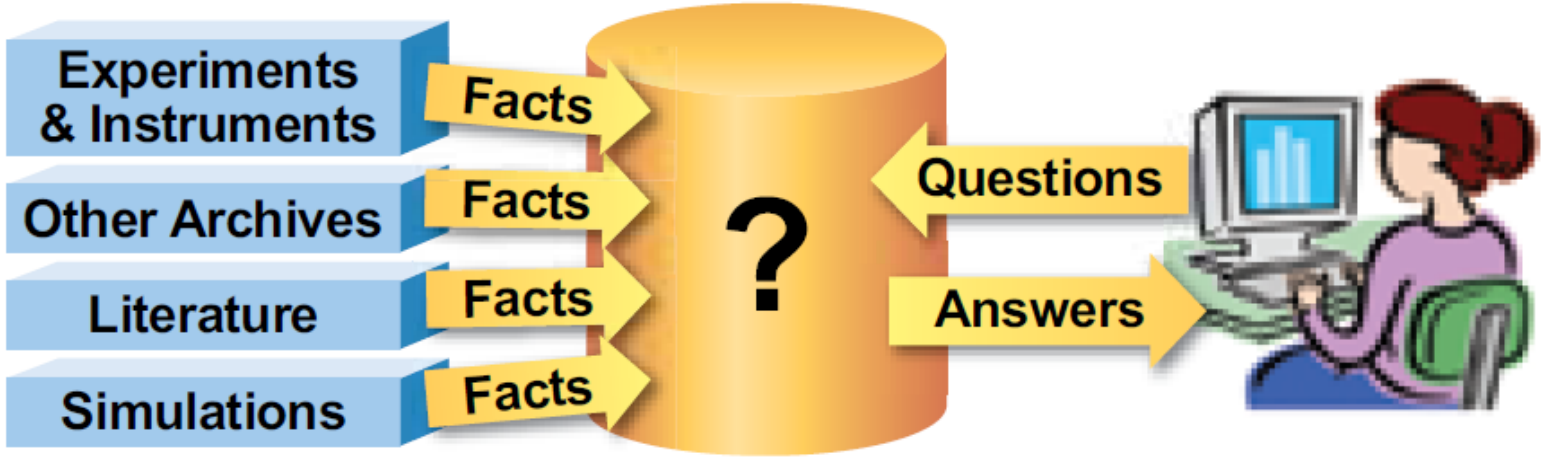
Famous Quotes

The advent of computation can be compared, in terms of the breadth and depth of its impact on research and scholarship, to the invention of writing and the development of modern mathematics.

Ian Foster, 2006

X-Info

- The evolution of X-Info and Comp-X for each discipline X
- How to codify and represent our knowledge



The Generic Problems

- Data ingest
- Managing a petabyte
- Common schema
- How to organize it
- How to reorganize it
- How to share it with others
- Query and Vis tools
- Building and executing models
- Integrating data and literature
- Documenting experiments
- Curation and long-term preservation

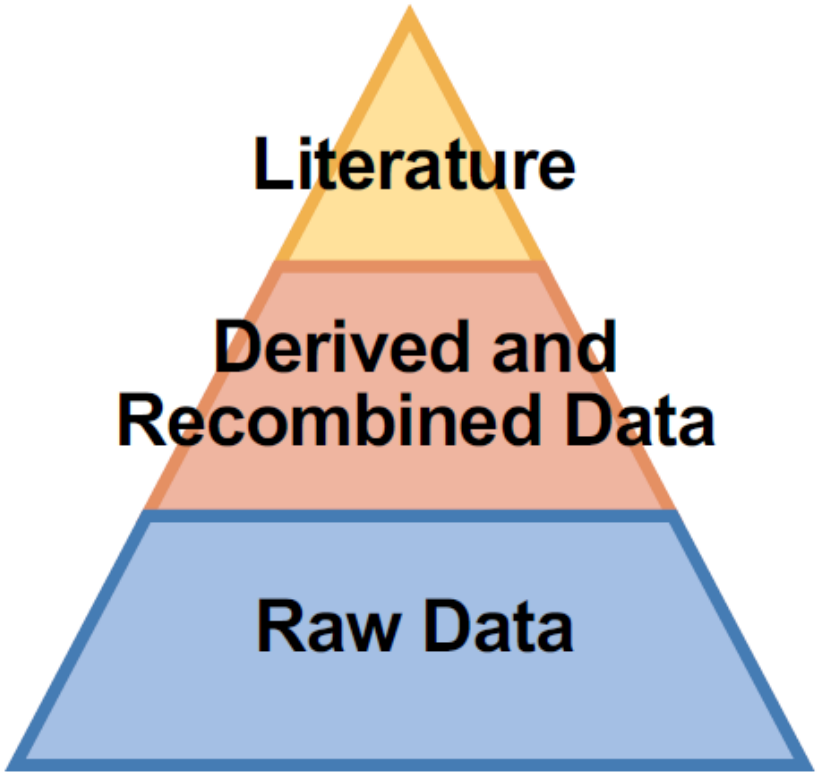
Famous Quotes

Computational thinking will be a fundamental skill used by everyone in the world by the middle of the 21st Century.

Jeanette Wing, 2006

All Scientific Data Online

- Many disciplines overlap and use data from other sciences
- Internet can unify all literature and data
- Go from literature to computation to data back to literature
- Information at your fingertips for everyone-everywhere



- Increase Scientific Information Velocity
- Huge increase in Science Productivity

Famous Quotes

The users should be able to focus their attention on the information content of the data, rather than how to discover, access, and use it.

Climate Change Science Program report, 2003

Summary

- Everything about science is changing because of the impact of information technology
- Experimental, theoretical, and computational science are all being affected by the data deluge, and a fourth, “data-intensive” science paradigm is emerging.
- Goal
  - A world in which all of the science literature is online
  - all of the science data is online
  - They interoperate with each other
- Lots of new tools are needed to make this happen.

Famous Quotes

A supercomputer is a device for turning compute-bound problems into I/O-bound problems.

Seymour Cray

Questions